

# 移动社交大数据

之  
微信朋友圈数据集、  
网络应用以及未来展望

姓名：李真晶

学号：SA19006031

参考文献：Zhang Y , Li Z , Gao C , et al. Mobile Social Big Data: WeChat Moments Dataset, Network Applications, and Opportunities[J]. IEEE Network, 2017, PP(99).

# 介绍

随着各种各样移动技术的高速发展，人们每时每刻都在产生大量新的社交数据。无论是对政府、企业还是研究机构而言，利用好这些数据都是非常重要的。

相比于基于web的社交网络，移动社交网络（MSN）有着更好的潜力与功能。如MSN能够提供基于位置的服务、移动通信和增强现实等功能。

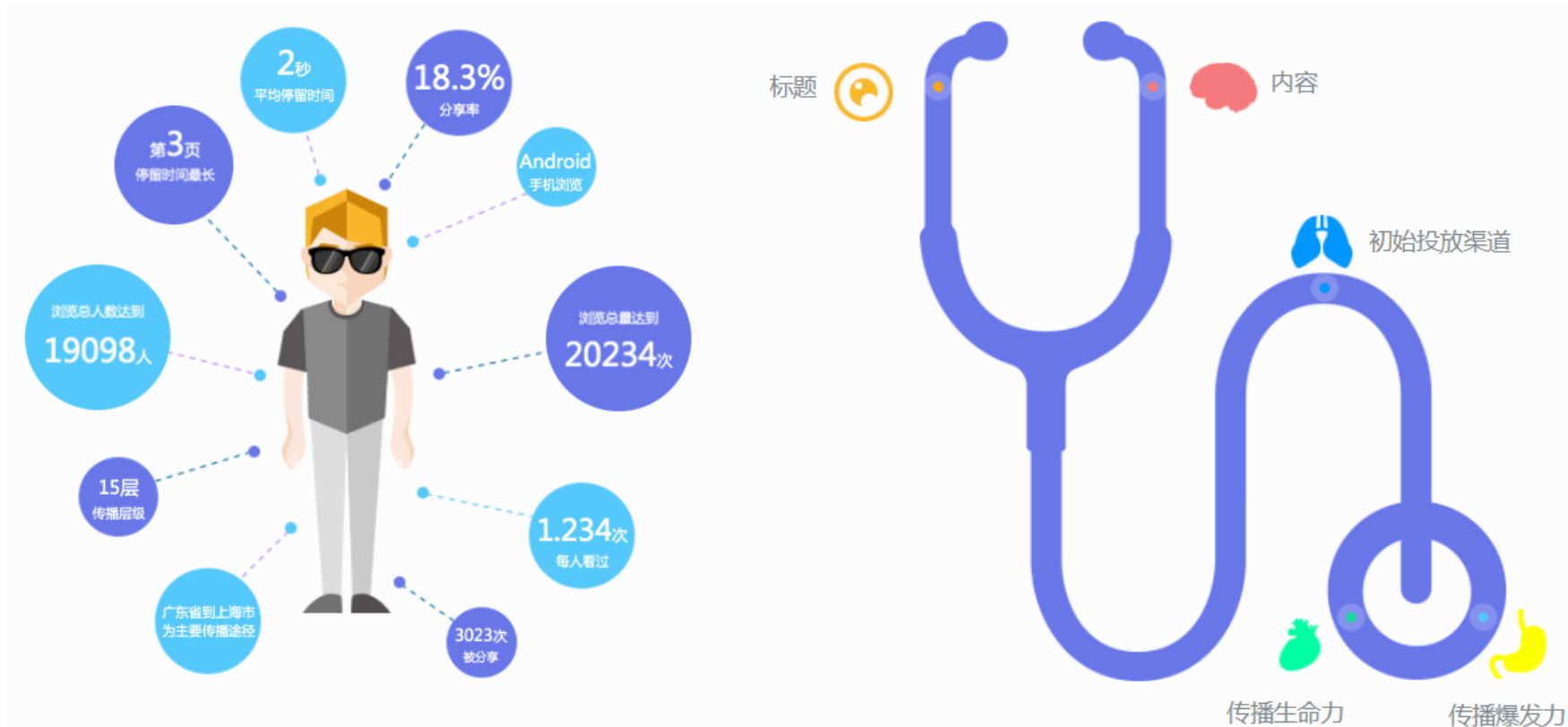
本文作者收集了一个微信朋友圈数据集——WeChatNet，并分析了三个基于WeChatNet的网络应用以及对未来的展望

# 主要内容

- 一、微信朋友圈数据集的收集
- 二、移动社交网络应用的分类
- 三、微信的新特性
- 四、基于数据分析下的网络应用
  - 1、移动蜂窝网络下的信息传播
  - 2、骨干网络流量预测
  - 3、流动人口分布预测
- 五、未来展望

# 微信朋友圈数据集的收集

使用 FIBODATA 获取微信页面的扩散轨迹。



- 从2016年1月14号到2016年2月17号，总共收集了 320,000 个页面传播轨迹，总共涉及到 25,133,330 名用户，246,369,415 条转发记录。
- 收集到的数据的格式为五元组 (5-tuple)

$\langle U_1, U_2, PID, IP, t \rangle$

$U_1$       该用户的帖子被看到

$U_2$       该用户看了U1的帖子

$PID$       帖子的ID

$IP$       U2的IP地址

$t$       这个“浏览帖子”事件的发生时间

移动社交网络(MSN)应用的对比

Service category	Service name	Twitter	Weibo	Facebook	WeChat	WhatsApp
Social networking services	Video sharing	✓	✓	✓	✓	✗
	Personal page	✓	✓	✓	✓	✗
	Post search	✓	✓	✓	✓	✗
	Favorite post save	✗	✗	✓	✓	✗
	Access to pages of non-friends	✓	✓	✓	✗	n/a
Messenger services	Limit on # of followers/friends	No limit	No limit	5000	5000	No limit
	Video/audio chat	✗	✗	✓	✓	✓
	Group messaging	✗	✓	✓	✓	✓
	Sending location	✗	✗	✓	✓	✓
	Voice messaging	✗	✗	✓	✓	✓
Miscellaneous services	Mobile payment	✗	✗	✓	✓	✗
	Video games	✗	✓	✓	✓	✗
	Offline services (taxi, ticket, etc.)	✗	✗	✗	✓	✗
	Shopping	✗	✓	✓	✓	✗

Twitter、weibo  
主要作为社交  
网络

WhatsApp 主要  
是即时通信

Facebook、  
WeChat  
结合了  
社交网络和  
即时通信

# 微信朋友圈（WM）的新特性

- 1、拥有强社交关系——在朋友圈中禁止进入陌生人的主页
- 2、选定内容显示——私密内容只会展示给选中的朋友
- 3、群聊——一种接触陌生人的方法



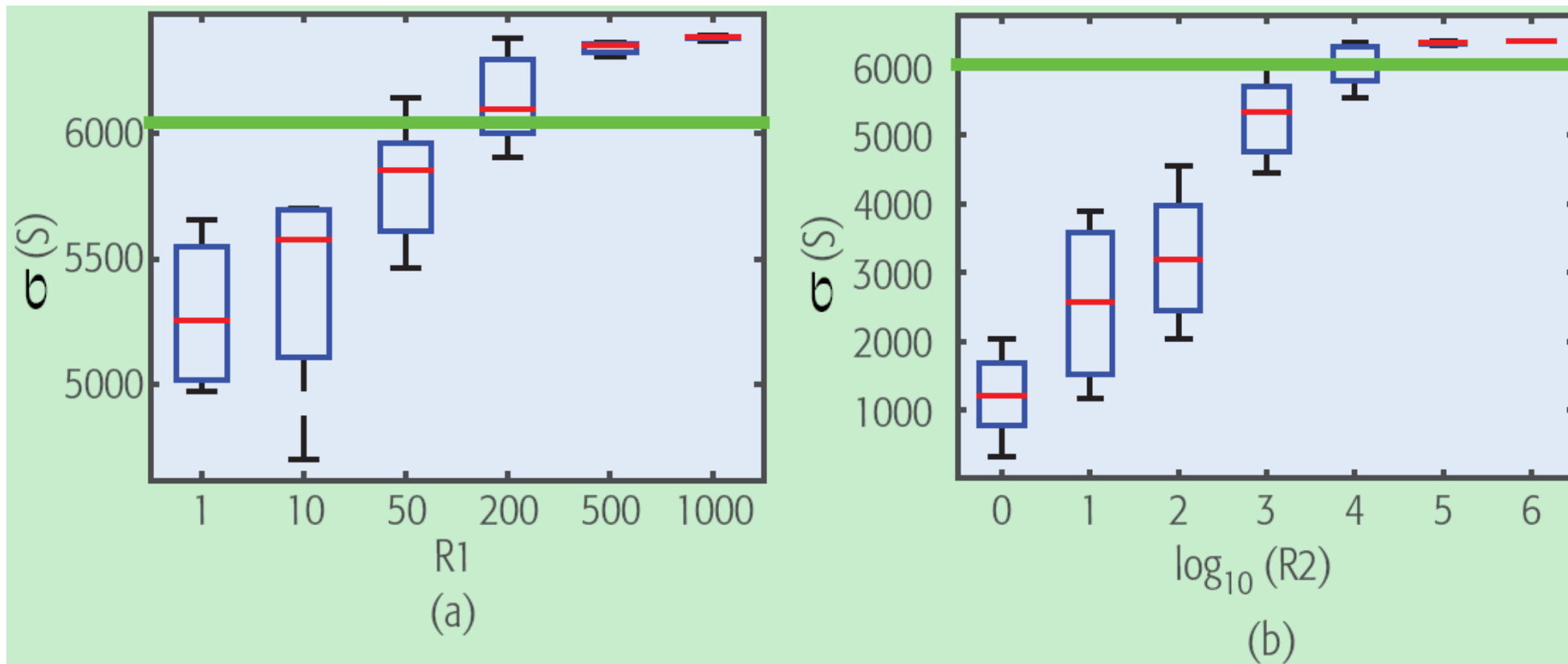
# 移动蜂窝网络下的信息传播

- 1、移动蜂窝网络中影响力最大的用户
  - =》移动社交网络中关键的意见领袖（KOLs，拥有大量的朋友或跟随者），他们能够有效地促进网络营销/广告传播等。
- 2、在WM中，选取KOL面临的挑战
- 有最大朋友数量的限制，一个名人或许在微信中也只有几百个朋友。

# 基于投票的策略 (voting-based strategy)

- 通过分析用户对信息扩散过程中的局部贡献来选择关键的结点 (KOLs)
- (1) 确定信息扩散过程中关键的扩散树
- (2) 在每一个扩散树中, 后代结点可以为其父节点投票 (微信网页是从父节点传播到子节点)
- 通过吉布斯采样 (Gibbs Sampling) 解决在WM中树的个数、结点很多的情况

# 实验结果

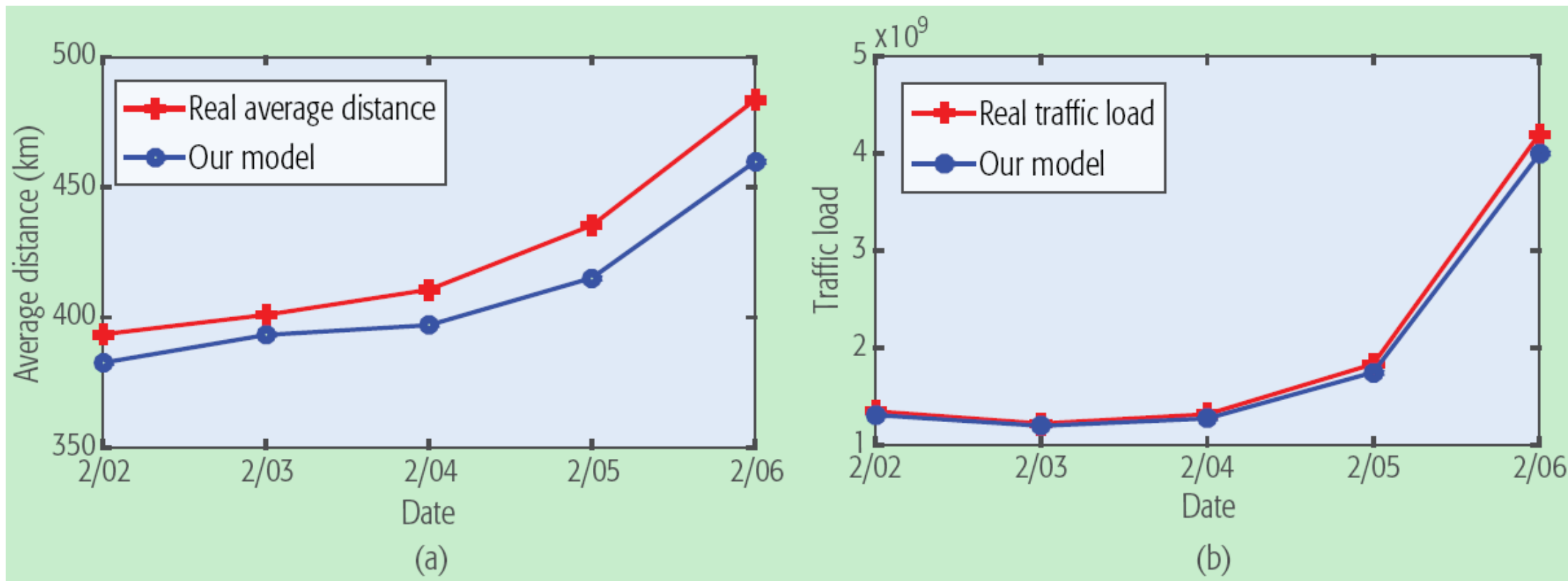


由 (a) 图可以看出, 随着 $R1$ 的增加,  $\delta(s)$ 的方差逐渐降低并达到一个稳定值, 当 $R1$ 超过200的时候, 投票策略优于贪婪算法。由 (b) 图可以看出,  $\delta(s)$ 随着 $R2$ 的增加而增加, 当 $R2$ 大于4的时候, 投票策略优于贪婪算法。

# 主干网络中的流量分布

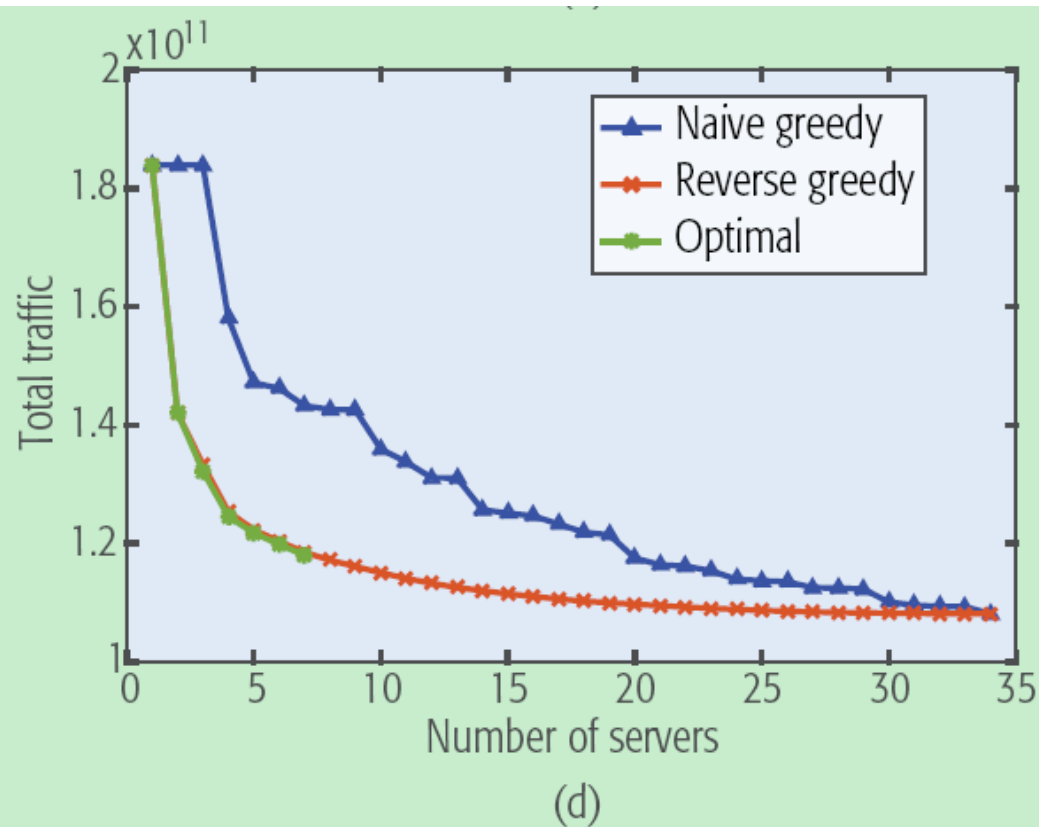
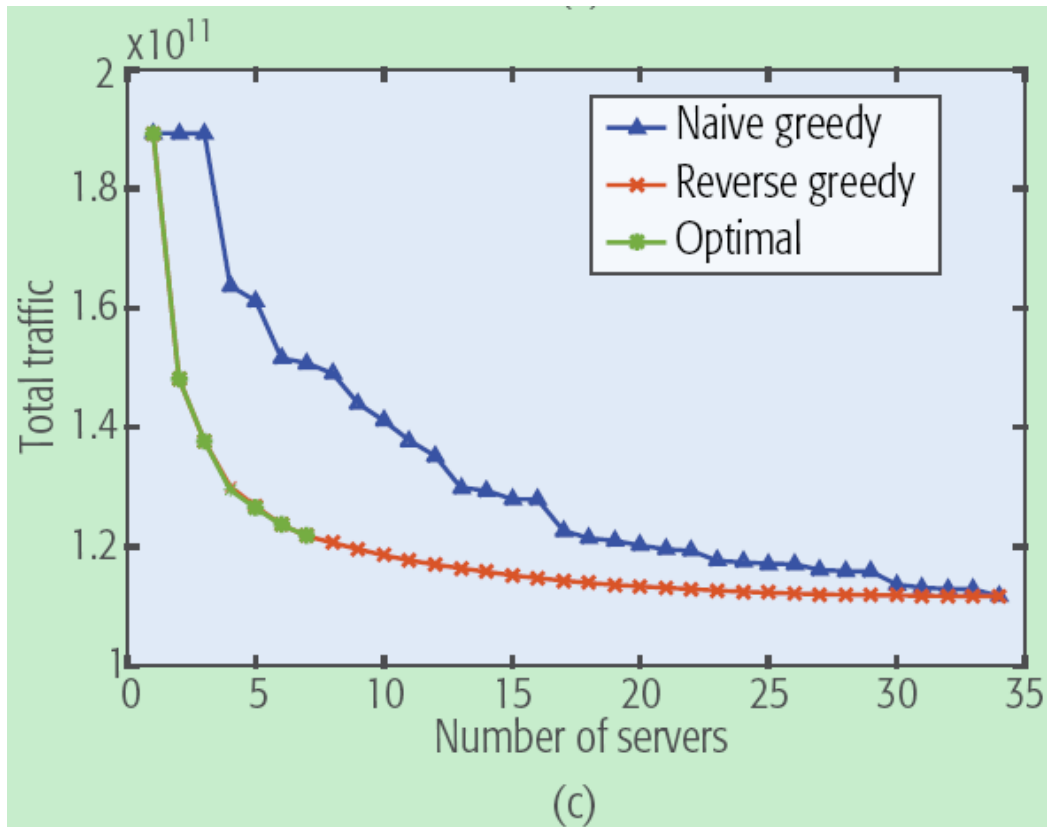
- 移动社交网络消息可以部分反映网络流量。通过在时间和空间维度上观察微信朋友之间的交互信息，去预测底层网络之间的流量负载，并提出了一个反向贪婪策略去放置服务器，平衡网络中的流量负载。

# 从通信模式到流量预测



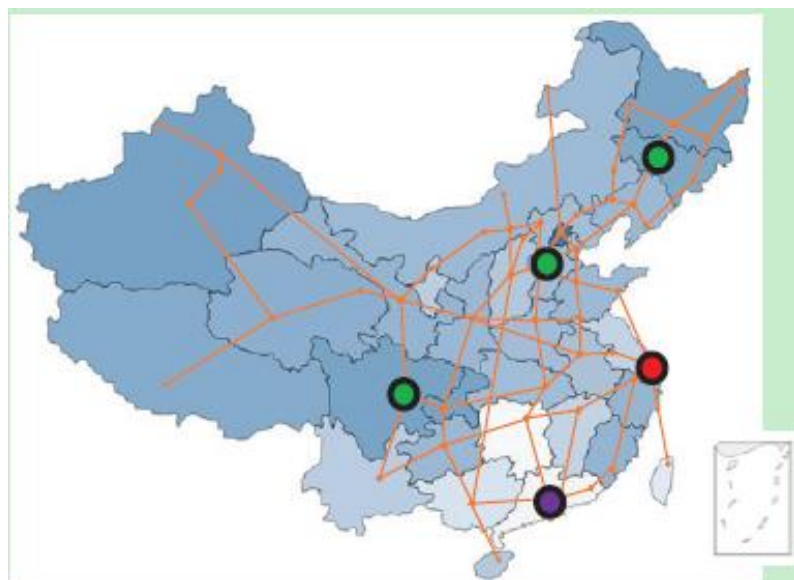
# 流量优化与服务器放置

- 目标：在合适的地方放置服务器，缩短用户之间的通信距离



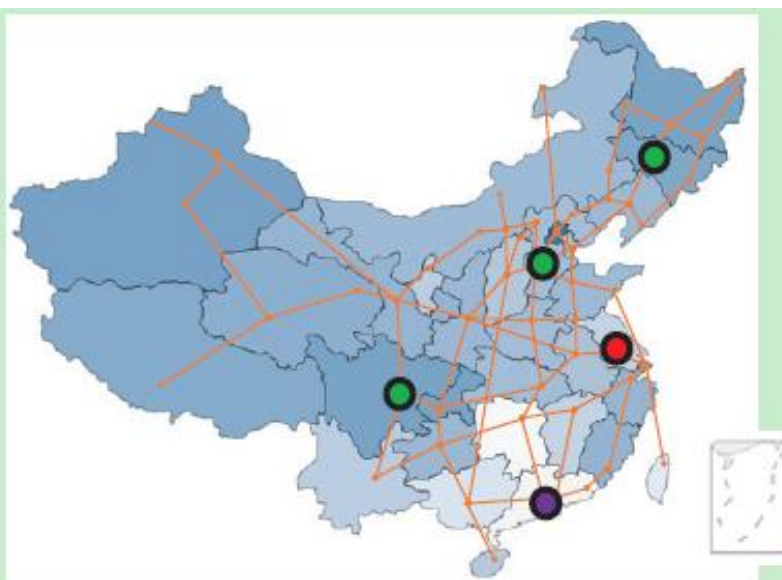
放置5台服务器后曲线变化平缓

# 放置5台服务器时的方案展示



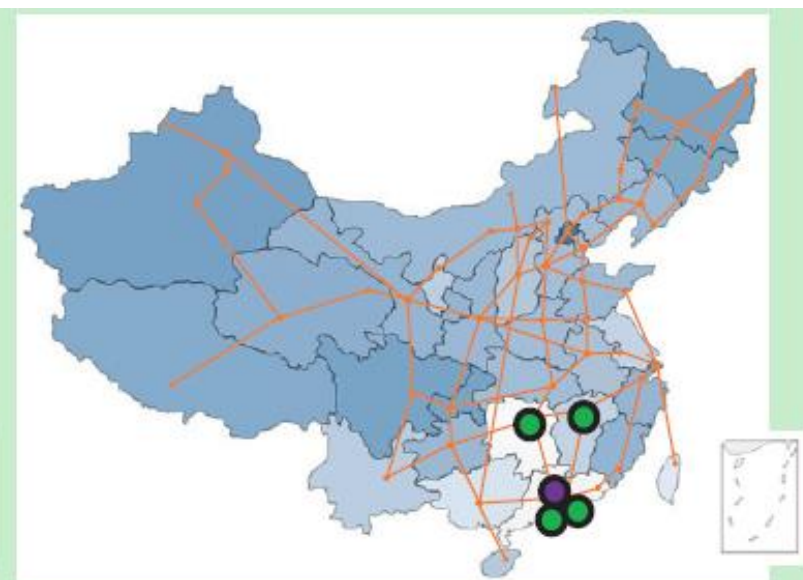
(e)

(a) 最优方案



(f)

(b) 反向贪婪策略



(g)

(g) 朴素贪婪算法

# 流动人口分布预测

- 预测人口在地理区域的分布对很多应用来说都很重要。
- 如在人口密度大的地方开展营销活动或加强公共安全。
- 传统方法通过线下的方法收集人口的流动轨迹，这在成本和数据可信度上都有很大不足。
- 智能手机上的移动社交软件的数据提供了更加的高效的方法。

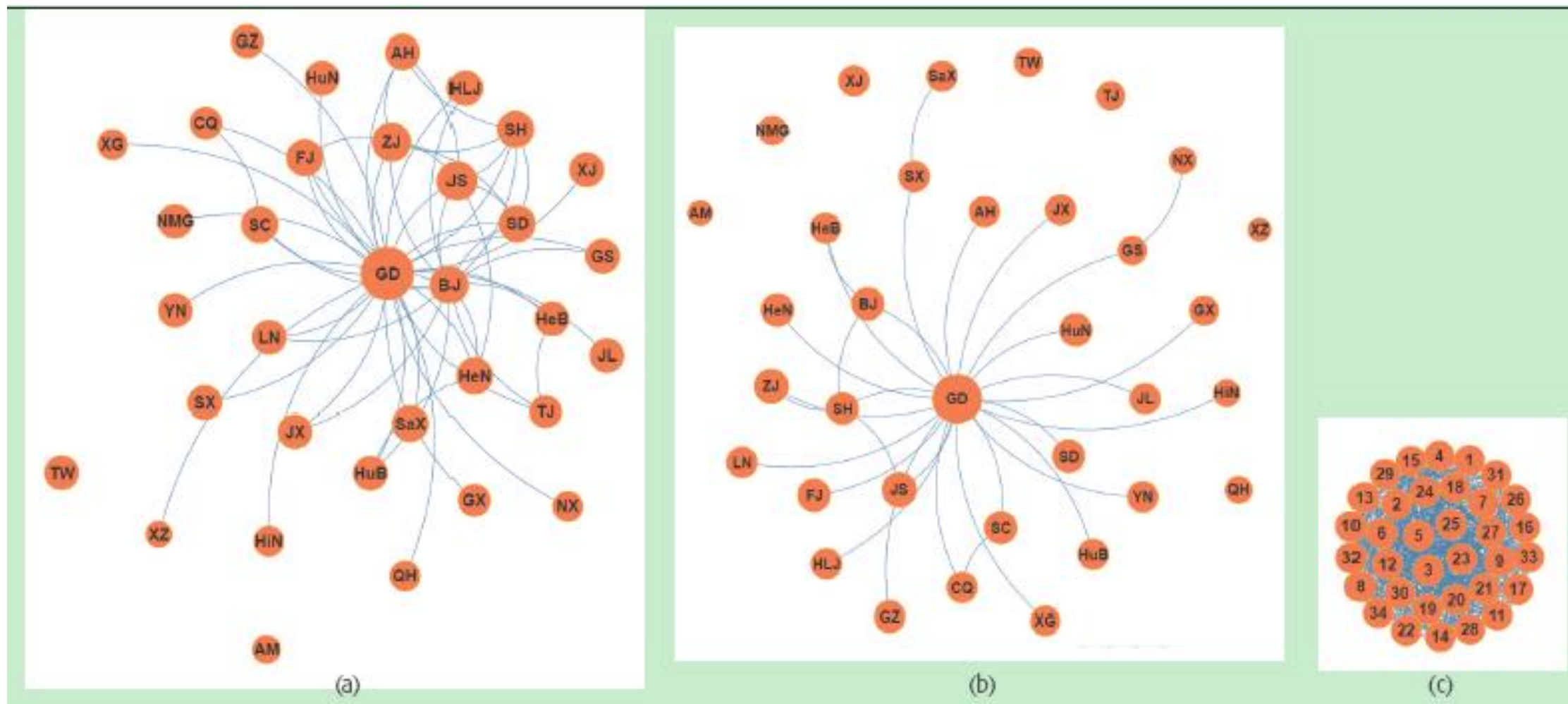


# 地理同质性建模

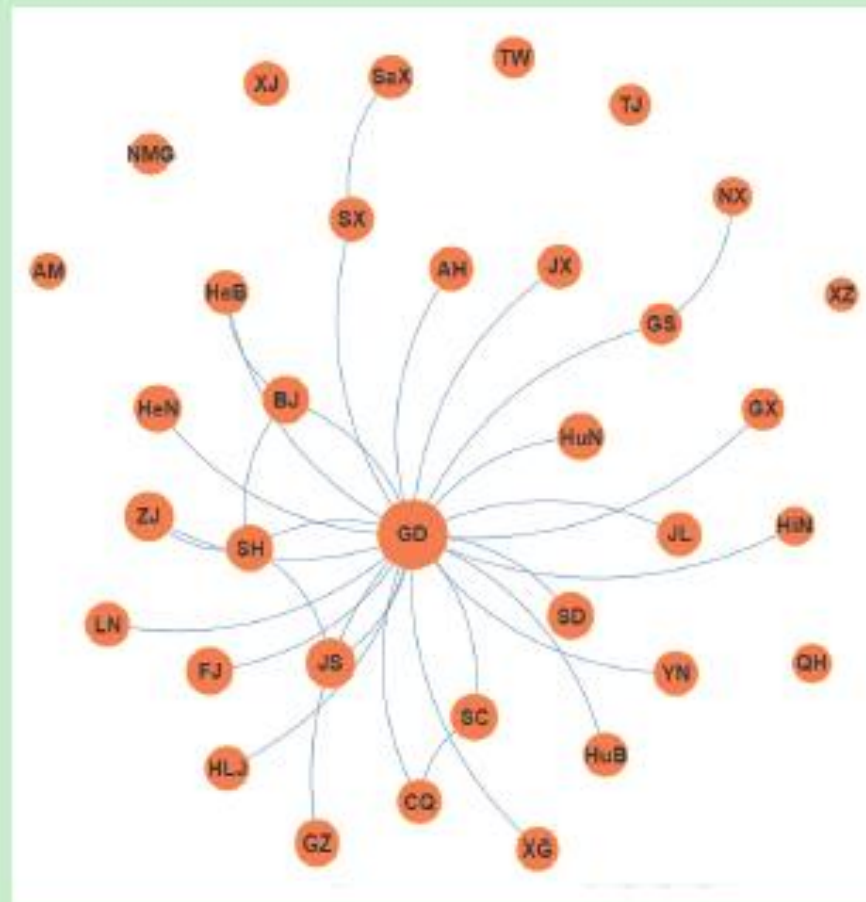
- 地理同质性（geo-homophily）：每一区域的人更倾向于和同一区域的人交流，而不是和其他区域的人。
- 我们可以通过移动社交网络用户之间的信息的地理位置来得到给定区域内的用户分布。基于这些分布进一步推断区域间的流动人口（（Dirichlet Process Mixture））。

# 微信朋友圈中的地理同质性

- 以中国34个省级行政区作为实验中的地理区域
- 在两个时间段分析了信息传播过程
  - (1) 春节前 (2016年1月14 - 2016年1月31)
  - (2) 春节



(a) 春节前



(b) 春节



(c) 基准网络

圆越大，省内信息传播越多；不同圆之间弧线越长，相应的省份间信息传播越多。（上图表明大部分的信息传播都在省内，台北和西藏之间甚至没有信息交流）

在春节的时候（图b），大部分的人都待在家，信息扩散主要集中在同一区域的朋友之间，因此省内的信息扩散增加，一些省间的信息扩散消失。

图c是一个无序的、没有地理同质性的对比图。圆与圆之间的弧线很密，大部分圆的大小都差不多。

# 未来展望

- (1) 探索新型移动社交网络（如WM）的结构，为什么它和其他的社交网络不同
- (2) 基于动态扩散图的营销
- (3) 深入地了解更多人口统计结果
- (4) 垃圾邮检测问题
- (5) 隐私泄露问题
- (6) 推广线下营销活动